

# Hessian Matrices and Statistics

1

## Example

$$f(x_1, x_2) = 4x_1^2e^{x_2^3}.$$

$$\implies \frac{\partial}{\partial x_1}f(x_1, x_2) = 8x_1e^{x_2^3}$$

$$\implies \frac{\partial^2}{\partial x_1^2}f(x_1, x_2) = 8e^{x_2^3}$$

$$\frac{\partial}{\partial x_2}f(x_1, x_2) = 12x_1^2x_2^2e^{x_2^3}$$

$$\implies \frac{\partial^2}{\partial x_1\partial x_2}f(x_1, x_2) = 24x_1x_2^2e^{x_2^3}$$

3

## Definition of Hessian Matrix

$$f : \mathbb{R}^d \rightarrow \mathbb{R}$$

$Hf(\mathbf{x})$  is the  $d \times d$  matrix with  $(i, j)$  entry

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}).$$

2

## Example (continued)

Eventually:

$$Hf(x_1, x_2) = \begin{bmatrix} 8e^{x_2^3} & 24x_1x_2^2e^{x_2^3} \\ 24x_1x_2^2e^{x_2^3} & 12x_1^2x_2(2 + 3x_2^3)e^{x_2^3} \end{bmatrix}$$

4

## Why Are Hessian Matrices Important in Statistics?

1. Fisher information of multi-parameter models.
2. Solution of multivariate optimisation problems

(e.g. maximum likelihood) via Newton-Raphson iteration.

5

6

## Fisher Information Result

Model is:

$$X_1, \dots, X_n \overset{\text{ind.}}{\sim} f_X(x; \boldsymbol{\theta})$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$  is a vector.

The Fisher information matrix is

$$I_n(\boldsymbol{\theta}) = E\{-H\ell(\boldsymbol{\theta})\}$$

where  $\ell(\boldsymbol{\theta})$  is the log-likelihood.

7

8

## Why Do We Care About Fisher Information?

If  $\hat{\theta}_j$  is the maximum likelihood estimate of  $\theta$  then the

asymptotic **standard error**

of  $\hat{\theta}_j$  is

$$\text{se}(\hat{\theta}_j) = \sqrt{[I_n(\boldsymbol{\theta})^{-1}]_{jj}}.$$

9

## Univariate Newton-Raphson Method

Goal: Maximise  $S(x)$  over  $x \in \mathbb{R}$ ;  $S$  is 'nicely behaved'.

Maximum occurs where  $S'(x) = 0$ .

$x_0$  = initial guess at solution.

$$x_{i+1} = x_i - S'(x_i)/S''(x_i).$$

$x_0, x_1, x_2, \dots$  converges (hopefully) to maximiser.

11

## Multivariate Newton-Raphson Method

Goal: Maximise  $S(\boldsymbol{x})$  over  $\boldsymbol{x} \in \mathbb{R}^d$ ;  $S$  is 'nicely behaved'.

Maximum occurs where  $DS(\boldsymbol{x}) = 0$ .

$\boldsymbol{x}_0$  = initial guess at solution.

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \{HS(\boldsymbol{x}_i)\}^{-1}\{DS(\boldsymbol{x}_i)\}^T.$$

$\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \dots$  converges (hopefully) to maximiser.

10

12

## Obtaining Hessian Matrices

For small multiparameter models such as  $N(\mu, \sigma^2)$  we can do directly using partial differentiation rules.

How about models such as the linear regression one:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})?$$

13

## Illustration

For the linear regression model can show:

$$d\ell(\boldsymbol{\beta}) = (1/\sigma^2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{X} d\boldsymbol{\beta}.$$

The second differential is:

15

## Second Identification Theorem of Vector Calculus

$$f : \mathbb{R}^d \rightarrow \mathbb{R}.$$

If  $\mathbf{A}$  is the  $d \times d$  matrix for which

$$d^2 f(\mathbf{x}) = (d\mathbf{x})^T \mathbf{A} (d\mathbf{x})$$

then

$$\mathbf{A} = \mathbf{H} f(\mathbf{x}).$$

14

$$\begin{aligned} d^2 \ell(\boldsymbol{\beta}) &= d\{(1/\sigma^2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{X} d\boldsymbol{\beta}\} \\ &= (1/\sigma^2)(-\mathbf{X} d\boldsymbol{\beta})^T \mathbf{X} d\boldsymbol{\beta} \\ &= (d\boldsymbol{\beta})^T (1/\sigma^2)(-\mathbf{X}^T \mathbf{X}) d\boldsymbol{\beta} \end{aligned}$$

By the **Second Identification Theorem**

$$\mathbf{H}\ell(\boldsymbol{\beta}) = -(1/\sigma^2)\mathbf{X}^T \mathbf{X}.$$

16

The Fisher information matrix is then (for  $\sigma$  known)

$$I_n(\boldsymbol{\beta}) = (1/\sigma^2)\mathbf{X}^T\mathbf{X}.$$